

On the Entropy of Couplings

Mladen Kovačević, *Student Member, IEEE*, Ivan Stanojević, *Member, IEEE*, and Vojin Šenk, *Member, IEEE*

Abstract—This paper studies bivariate distributions with fixed marginals from an information-theoretic perspective. In particular, continuity and related properties of various information measures (Shannon entropy, conditional entropy, mutual information, Rényi entropy) on the set of all such distributions are investigated. The notion of minimum entropy coupling is introduced, and it is shown that it defines a family of (pseudo)metrics on the space of all probability distributions in the same way as the so-called maximal coupling defines the total variation distance. Some basic properties of these pseudometrics are established, in particular their relation to the total variation distance, and a new characterization of the conditional entropy is given. Finally, some natural optimization problems associated with the above information measures are identified and shown to be NP-hard. Their special cases are found to be essentially information-theoretic restatements of well-known computational problems, such as the SUBSET SUM and PARTITION problems.

Index Terms—Coupling, distributions with fixed marginals, information measures, Rényi entropy, continuity of entropy, entropy minimization, maximization of mutual information, subset sum, entropy metric, infinite alphabet, measure of dependence.

I. INTRODUCTION

DISTRIBUTIONS with fixed marginals have been studied extensively in the probability literature (see for example [32] and the references therein). They are closely related to (and sometimes identified with, as will be the case in this paper) the concept of coupling, which has proven to be a very useful proof technique in probability theory [35], and in particular in the theory of Markov chains [25]. There is also rich literature on the geometrical and combinatorial properties of sets of distributions with given marginals, which are known as transportation polytopes in this context (see, e.g., [7]). We investigate here these objects from a certain information-theoretic perspective. Our results and the general outline of the paper are briefly described below.

Section II provides definitions and elementary properties of the functionals studied subsequently – Shannon entropy, Rényi entropy, conditional entropy, mutual information, and information divergence. In Section III we recall the definition and basic properties of couplings, i.e., bivariate distributions with fixed marginals, and introduce the corresponding notation. The notion of minimum entropy coupling, which will be useful in subsequent analysis, is also introduced here. In Section IV we discuss in detail continuity and related properties of the

above-mentioned information measures under constraints on the marginal distributions. These results complement rich literature on the topic of extending the statements of information theory to the case of countably infinite alphabets.

In Section V we define a family of (pseudo)metrics on the space of probability distributions, that is based on the minimum entropy coupling in the same way as the total variation distance is based on the so-called maximal coupling. The relation between these distances is derived from the Fano's inequality. Some other properties of the new metrics are also discussed, in particular an interesting characterization of the conditional entropy that they yield.

In Section VI certain optimization problems associated with the above-mentioned information measures are studied. Most of them are, in a certain sense, the reverse problems of the well-known optimization problems, such as the maximum entropy principle, the channel capacity, and the information projections. The general problems of (Rényi) entropy minimization, maximization of mutual information, and maximization of information divergence are all shown to be intractable. Since mutual information is a good measure of dependence of two random variables, this will also lead to a similar result for all measures of dependence satisfying Rényi's axioms, and to a statistical scenario where this result might be of interest. The potential practical relevance of these problems is also discussed in this section, as well as their theoretical value. Namely, all of them are found to be basically restatements of some well-known problems in complexity theory.

II. INFORMATION MEASURES

In this introductory section we recall the definitions and elementary properties of some basic information-theoretic functionals. All random variables are assumed to be discrete, with alphabet \mathbb{N} – the set of positive integers, or a subset of \mathbb{N} of the form $\{1, \dots, n\}$.

Shannon entropy of a random variable X with probability distribution $P = (p_i)$ (we also sometimes write $P(i)$ for the masses of P) is defined as:

$$H(X) \equiv H(P) = - \sum_i p_i \log p_i \quad (1)$$

with the usual convention $0 \log 0 = 0$ being understood. The base of the logarithm, $b > 1$, is arbitrary and will not be specified. H is a strictly concave¹ functional in P [9]. Further, for a pair of random variables (X, Y) with joint distribution $S = (s_{i,j})$ and marginal distributions $P = (p_i)$ and $Q = (q_j)$, the following defines their joint entropy:

$$H(X, Y) \equiv H_{X,Y}(S) = - \sum_{i,j} s_{i,j} \log s_{i,j}, \quad (2)$$

¹ To avoid possible confusion concave means \cap and convex means \cup .

Date: March 13, 2013.

This work was supported by the Ministry of Science and Technological Development of the Republic of Serbia (grants No. TR32040 and III44003). Part of the work was presented at the 2012 IEEE Information Theory Workshop (ITW).

The authors are with the Department of Electrical Engineering, Faculty of Technical Sciences, University of Novi Sad, Serbia. E-mails: {kmladen, cet_ivan, vojnin_senk}@uns.ac.rs.

conditional entropy:

$$H(X|Y) \equiv H_{X|Y}(S) = - \sum_{i,j} s_{i,j} \log \frac{s_{i,j}}{q_j}, \quad (3)$$

and mutual information:

$$I(X;Y) \equiv I_{X;Y}(S) = \sum_{i,j} s_{i,j} \log \frac{s_{i,j}}{p_i q_j}, \quad (4)$$

again with appropriate conventions. We will refer to the above quantities as the Shannon information measures. They are all related by simple identities:

$$\begin{aligned} H(X,Y) &= H(X) + H(Y) - I(X;Y) \\ &= H(X) + H(Y|X) \end{aligned} \quad (5)$$

and obey the following inequalities:

$$\max \{H(X), H(Y)\} \leq H(X,Y) \leq H(X) + H(Y), \quad (6)$$

$$\min \{H(X), H(Y)\} \geq I(X;Y) \geq 0, \quad (7)$$

$$0 \leq H(X|Y) \leq H(X). \quad (8)$$

The equalities on the right-hand sides of (6)–(8) are achieved if and only if X and Y are independent. The equalities on the left-hand sides of (6) and (7) are achieved if and only if X deterministically depends on Y (i.e., iff X is a function of Y), or vice versa. The equality on the left-hand side of (8) holds if and only if X deterministically depends on Y . We will use some of these properties in our proofs; for their demonstration we point the reader to the standard reference [9].

From identities (5) one immediately observes the following: Over a set of bivariate probability distributions with fixed marginals (and hence fixed marginal entropies $H(X)$ and $H(Y)$), all the above functionals differ up to an additive constant (and a minus sign in the case of mutual information), and hence one can focus on studying only one of them and easily translate the results for the others. This fact will also be exploited later.

Relative entropy (information divergence, Kullback-Leibler divergence) $D(P||Q)$ is the following functional:

$$D(P||Q) = \sum_i p_i \log \frac{p_i}{q_i}. \quad (9)$$

Finally, Rényi entropy [29] of order $\alpha \geq 0$ of a random variable X with distribution P is defined as:

$$H_\alpha(X) \equiv H_\alpha(P) = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha, \quad (10)$$

with

$$H_0(P) = \lim_{\alpha \rightarrow 0} H_\alpha(P) = \log |P| \quad (11)$$

where $|P| = |\{i : p_i > 0\}|$ denotes the size of the support of P , and

$$H_1(P) = \lim_{\alpha \rightarrow 1^+} H_\alpha(P) = H(P). \quad (12)$$

One can also define:

$$H_\infty(P) = \lim_{\alpha \rightarrow \infty} H_\alpha(P) = -\log \max_i p_i. \quad (13)$$

Joint Rényi entropy of the pair (X, Y) having distribution $S = (s_{i,j})$ is naturally defined as:

$$H_\alpha(X, Y) \equiv H_\alpha(S) = \frac{1}{1-\alpha} \log \sum_{i,j} s_{i,j}^\alpha. \quad (14)$$

By using subadditivity (for $\alpha < 1$) and superadditivity (for $\alpha > 1$) properties of the function x^α one concludes that:

$$H_\alpha(X, Y) \geq \max \{H_\alpha(X), H_\alpha(Y)\} \quad (15)$$

with equality if and only if X is a function of Y , or vice versa. However, Rényi analogue of the right-hand side of (6) does not hold unless $\alpha = 0$ or $\alpha = 1$ [1]. In fact, no upper bound on the joint Rényi entropy in terms of the marginal entropies can exist for $0 < \alpha < 1$, as will be illustrated in Section IV.

III. COUPLINGS OF PROBABILITY DISTRIBUTIONS

A coupling of two probability distributions P and Q is a bivariate distribution S (on the product space, in our case \mathbb{N}^2) with marginals P and Q . This concept can also be defined for random variables in a similar manner, and it represents a powerful proof technique in probability theory [35].

Let $\Gamma_n^{(1)}$ and $\Gamma_{n \times m}^{(2)}$ denote the sets of one- and two-dimensional probability distributions with alphabets of size n and $n \times m$, respectively:

$$\Gamma_n^{(1)} = \left\{ (p_i) \in \mathbb{R}^n : p_i \geq 0, \sum_i p_i = 1 \right\} \quad (16)$$

$$\Gamma_{n \times m}^{(2)} = \left\{ (p_{i,j}) \in \mathbb{R}^{n \times m} : p_{i,j} \geq 0, \sum_{i,j} p_{i,j} = 1 \right\} \quad (17)$$

and let $\mathcal{C}(P, Q)$ denote the set of all couplings of $P \in \Gamma_n^{(1)}$ and $Q \in \Gamma_m^{(1)}$:

$$\mathcal{C}(P, Q) = \left\{ S \in \Gamma_{n \times m}^{(2)} : \sum_j s_{i,j} = p_i, \sum_i s_{i,j} = q_j \right\}. \quad (18)$$

It is easy to show that the sets $\mathcal{C}(P, Q)$ are convex and closed in $\Gamma_{n \times m}^{(2)}$. They are also clearly disjoint and cover entire $\Gamma_{n \times m}^{(2)}$, i.e., they form a partition of $\Gamma_{n \times m}^{(2)}$. Finally, they are parallel affine $(n-1)(m-1)$ -dimensional subspaces of the $(n \cdot m - 1)$ -dimensional space $\Gamma_{n \times m}^{(2)}$. (We have in mind the restriction of the corresponding affine spaces in $\mathbb{R}^{n \times m}$ to $\mathbb{R}_+^{n \times m}$.)

The set of distributions with fixed marginals is basically the set of matrices with nonnegative entries and prescribed row and column sums (only now the total sum is required to be one, but this is inessential). Such sets are special cases of the so-called transportation polytopes [7].

We shall also find it interesting to study information measures over the sets of distributions whose one marginal and the support of the other are fixed:

$$\mathcal{C}(P, m) = \bigcup_{Q \in \Gamma_m^{(1)}} \mathcal{C}(P, Q). \quad (19)$$

These sets are also convex polytopes and form a partition of $\Gamma_{n \times m}^{(2)}$ for $P \in \Gamma_n^{(1)}$.

A. Minimum entropy couplings

We now introduce one special type of couplings which will be useful in subsequent analysis.

Definition 1: Minimum entropy coupling of probability distributions P and Q is a bivariate distribution $S^* \in \mathcal{C}(P, Q)$ which minimizes the entropy functional $H \equiv H_{X,Y}$, i.e.,

$$H(S^*) = \inf_{S \in \mathcal{C}(P, Q)} H(S). \quad (20)$$

Minimum entropy couplings exist for any $P \in \Gamma_n^{(1)}$ and $Q \in \Gamma_m^{(1)}$ because sets $\mathcal{C}(P, Q)$ are compact (closed and bounded) and entropy is continuous over $\Gamma_{n \times m}^{(2)}$ and hence attains its extrema. (Note, however, that they need not be unique.) From the strict concavity of entropy one concludes that the minimum entropy couplings must be vertices of $\mathcal{C}(P, Q)$ (i.e., they cannot be expressed as $aS + (1-a)T$, with $S, T \in \mathcal{C}(P, Q)$, $a \in (0, 1)$). Finally, from identities (5) it follows that the minimizers of $H_{X,Y}$ over $\mathcal{C}(P, Q)$ are simultaneously the minimizers of $H_{X|Y}$ and $H_{Y|X}$ and the maximizers of $I_{X,Y}$, and hence could also be called *maximum mutual information couplings* for example.

Definition 1 (cont.): Minimum α -entropy coupling of probability distributions P and Q is a bivariate distribution $S^* \in \mathcal{C}(P, Q)$ which minimizes the Rényi entropy functional H_α .

Similarly to the above, existence of the minimum α -entropy couplings is easy to establish, as is the fact that they must be vertices of $\mathcal{C}(P, Q)$ (H_α is concave for $0 \leq \alpha \leq 1$; for $\alpha > 1$ it is neither concave nor convex [4] but the claim follows from the convexity of $\sum_{i,j} s_{i,j}^\alpha$).

IV. INFINITE ALPHABETS

We now establish some basic properties of information measures over $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$, and of the sets $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$ themselves, in the case when the distributions P and Q have possibly infinite supports. The notation is similar to the finite alphabet case, for example:

$$\begin{aligned} \Gamma^{(1)} &= \left\{ (p_i)_{i \in \mathbb{N}} : p_i \geq 0, \sum_i p_i = 1 \right\}, \\ \Gamma^{(2)} &= \left\{ (p_{i,j})_{i,j \in \mathbb{N}} : p_{i,j} \geq 0, \sum_{i,j} p_{i,j} = 1 \right\}. \end{aligned} \quad (21)$$

The following well-known claim will be useful. We give a proof for completeness.

Lemma 2: Let $f : A \rightarrow \mathbb{R}$, with $A \subseteq \mathbb{R}$ closed, be a continuous nonnegative function. Then the functional $F(x) = \sum_i f(x_i)$, $x = (x_1, x_2, \dots)$, is lower semi-continuous in ℓ^1 topology.

Proof: Let $\|x^{(n)} - x\|_1 \rightarrow 0$. Then, by using nonnegativity and continuity of f , we obtain

$$\begin{aligned} \liminf_{n \rightarrow \infty} F(x^{(n)}) &= \liminf_{n \rightarrow \infty} \sum_{i=1}^{\infty} f(x_i^{(n)}) \\ &\geq \liminf_{n \rightarrow \infty} \sum_{i=1}^K f(x_i^{(n)}) \\ &= \sum_{i=1}^K f(x_i), \end{aligned} \quad (22)$$

where the fact that $\|x^{(n)} - x\|_1 \rightarrow 0$ implies $|x_i^{(n)} - x_i| \rightarrow 0$, $\forall i$, was also used. Letting $K \rightarrow \infty$ we get

$$\liminf_{n \rightarrow \infty} F(x^{(n)}) \geq F(x), \quad (23)$$

which was to be shown. ■

A. Compactness of $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$

Let $\ell_2^1 = \{(x_{i,j})_{i,j \in \mathbb{N}} : \sum_{i,j} |x_{i,j}| < \infty\}$. This is the familiar ℓ^1 space, only defined for two-dimensional sequences. It clearly shares all the essential properties of ℓ^1 , completeness being the one we shall exploit. The metric understood is:

$$\|x - y\|_1 = \sum_{i,j} |x_{i,j} - y_{i,j}|, \quad (24)$$

for $x, y \in \ell_2^1$. In the context of probability distributions, this distance is usually called the total variation distance (actually, it is twice the total variation distance, see (49)).

Theorem 3: For any $P, Q \in \Gamma^{(1)}$ and $m \in \mathbb{N}$, $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$ are compact.

Proof: A metric space is compact if and only if it is complete and totally bounded [27, Thm 45.1]. These facts are demonstrated in the following two propositions. ■

Proposition 4: $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$ are complete metric spaces.

Proof: It is enough to show that $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$ are closed in ℓ_2^1 because closed subsets of complete spaces are always complete [27]. In other words, it suffices to show that for any sequence $S_n \in \mathcal{C}(P, Q)$ converging to some $S \in \ell_2^1$ (in the sense that $\|S_n - S\|_1 \rightarrow 0$), we have $S \in \mathcal{C}(P, Q)$. This is straightforward. If S_n all have the same marginals (P and Q), then S must also have these marginals, for otherwise the distance between S_n and S would be lower bounded by the distance between the corresponding marginals:

$$\sum_{i,j} |S(i,j) - S_n(i,j)| \geq \sum_i \left| \sum_j S(i,j) - S_n(i,j) \right| \quad (25)$$

and hence could not decrease to zero. The case of $\mathcal{C}(P, m)$ is similar. ■

For our next claim, recall that a set E is said to be totally bounded if it has a finite covering by ϵ -balls, for any $\epsilon > 0$. In other words, for any $\epsilon > 0$, there exist $x_1, \dots, x_K \in E$ such that $E \subseteq \bigcup_k \mathcal{B}(x_k, \epsilon)$, where $\mathcal{B}(x_k, \epsilon)$ denotes the open ball around x_k of radius ϵ . The points x_1, \dots, x_K are then called an ϵ -net for E .

Proposition 5: $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$ are totally bounded.

Proof: We prove the statement for $\mathcal{C}(P, Q)$, the proof for $\mathcal{C}(P, m)$ is very similar. Let P, Q , and $\epsilon > 0$ be given. We need to show that there exist distributions $S_1, \dots, S_K \in \mathcal{C}(P, Q)$ such that $\mathcal{C}(P, Q) \subseteq \bigcup_k \mathcal{B}(S_k, \epsilon)$, and this is done in the following. There exists N such that $\sum_{i=N+1}^{\infty} p_i < \frac{\epsilon}{6}$ and $\sum_{j=N+1}^{\infty} q_j < \frac{\epsilon}{6}$. Observe the truncations of the distributions P and Q , namely (p_1, \dots, p_N) and (q_1, \dots, q_N) . Assume that $\sum_{i=1}^N p_i \geq \sum_{j=1}^N q_j$, and let $r = \sum_{i=1}^N p_i - \sum_{j=1}^N q_j$ (otherwise, just interchange P and Q). Now let $P^{(N)} = (p_1, \dots, p_N)$ and $Q^{(N,r)} = (q_1, \dots, q_N, r)$, and observe $\mathcal{C}(P^{(N)}, Q^{(N,r)})$. (Adding r was necessary for

$\mathcal{C}(P^{(N)}, Q^{(N,r)})$ to be nonempty.) This set is closed (see the proof of Proposition 4) and bounded in $\mathbb{R}^{N \times (N+1)}$, and hence it is compact by the Heine-Borel theorem. This further implies that it is totally bounded and has an $\frac{\epsilon}{6}$ -net, i.e., there exist $T_1, \dots, T_K \in \mathcal{C}(P^{(N)}, Q^{(N,r)})$ such that $\mathcal{C}(P^{(N)}, Q^{(N,r)}) \subseteq \bigcup_k \mathcal{B}(T_k, \frac{\epsilon}{6})$. Now construct distributions $S_1, \dots, S_K \in \mathcal{C}(P, Q)$ by “padding” T_1, \dots, T_K . Namely, take S_k to be any distribution in $\mathcal{C}(P, Q)$ which coincides with T_k on the first $N \times N$ coordinates, for example:

$$S_k(i, j) = \begin{cases} T_k(i, j), & i, j \leq N \\ 0, & j \leq N, i > N \\ T_k(i, N+1) \cdot q_j / \sum_{j=N+1}^{\infty} q_j, & i \leq N, j > N \\ p_i \cdot q_j / \sum_{j=N+1}^{\infty} q_j, & i, j > N. \end{cases} \quad (26)$$

Note that $\|T_\ell - S_\ell\|_1 < \frac{\epsilon}{3}$ (where we understand that $T_\ell(i, j) = 0$ for $i > N$ or $j > N+1$). We prove below that S_k ’s are the desired ϵ -net for $\mathcal{C}(P, Q)$, i.e., that any distribution $S \in \mathcal{C}(P, Q)$ is at distance at most ϵ from some S_ℓ , $\ell \in \{1, \dots, K\}$ ($\|S - S_\ell\|_1 < \epsilon$). Observe some $S \in \mathcal{C}(P, Q)$, and let S' be its $N \times N$ truncation:

$$S'(i, j) = \begin{cases} S(i, j), & i, j \leq N \\ 0, & \text{otherwise.} \end{cases} \quad (27)$$

Note that S' is not a distribution, but that does not affect the proof. Note also that the marginals of S' are bounded from above by the marginals of S , namely $q'_j = \sum_i S'(i, j) \leq q_j$ and $p'_i = \sum_j S'(i, j) \leq p_i$. Finally, we have $\|S - S'\|_1 < \frac{\epsilon}{3}$ because the total mass of S on the coordinates where $i > N$ or $j > N$ is at most $\frac{\epsilon}{3}$. The next step is to create $S'' \in \mathcal{C}(P^{(N)}, Q^{(N,r)})$ by adding masses to S' on the $N \times (N+1)$ rectangle. One way to do this is as follows. Let

$$u_i = \begin{cases} p_i - p'_i, & i \leq N \\ 0, & i > N \end{cases}, \quad (28)$$

$$v_j = \begin{cases} q_j - q'_j, & j \leq N \\ r, & j = N+1 \\ 0, & j > N+1 \end{cases}, \quad (29)$$

and let $U = (u_i)$, and $V = (v_j)$, and $c = \sum_i u_i = \sum_j v_j$. Now define S'' by:

$$S'' = S' + \frac{1}{c} U \times V. \quad (30)$$

It is easy to verify that $S'' \in \mathcal{C}(P^{(N)}, Q^{(N,r)})$ and that $\|S' - S''\|_1 < \frac{\epsilon}{6}$ because the total mass added is

$$\begin{aligned} c &= \sum_{i=1}^N p_i - p'_i = \sum_{i=1}^N \sum_{j=1}^{\infty} (S(i, j) - S'(i, j)) \\ &= \sum_{i=1}^N \sum_{j=N+1}^{\infty} S(i, j) \\ &\leq \sum_{j=N+1}^{\infty} q_j < \frac{\epsilon}{6}. \end{aligned} \quad (31)$$

Now recall that T_k ’s form an $\frac{\epsilon}{6}$ -net for $\mathcal{C}(P^{(N)}, Q^{(N,r)})$ and consequently that there exists some T_ℓ , $\ell \in \{1, \dots, K\}$, with

$\|S'' - T_\ell\|_1 < \frac{\epsilon}{6}$. To put this all together, write:

$$\begin{aligned} \|S - S_\ell\|_1 &\leq \|S - S'\|_1 + \|S' - S''\|_1 + \\ &\quad \|S'' - T_\ell\|_1 + \|T_\ell - S_\ell\|_1 < \epsilon, \end{aligned} \quad (32)$$

which completes the proof. \blacksquare

B. Continuity of Shannon information measures

Shannon information measures are known to be discontinuous functionals in general [17], [39]. Imposing certain restrictions on the marginal distributions and entropies, however, ensures their continuity.

Theorem 6: Let $P, Q \in \Gamma^{(1)}$ and $m \in \mathbb{N}$, and assume that Q has finite entropy. Then Shannon information measures are continuous over $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$.

Proof: The claim can be established by using [14, Thm 4.3] and exhibiting the cost-stable codes for these statistical models, but we also give here a more direct proof (which will then be extended to prove Theorem 8). Write

$$H_Y(S) = I_{X,Y}(S) + H_{Y|X}(S). \quad (33)$$

The functional $H_{Y|X}(S) = \sum_{i,j} s_{i,j} \log \frac{p_i}{s_{i,j}}$ is lower semi-continuous by Lemma 2. The functional $I_{X,Y}$ is also lower semi-continuous since

$$I_{X,Y}(S) = D(S||P \times Q), \quad (34)$$

and information divergence $D(S||T)$ is known to be jointly lower semi-continuous in the distributions S and T [36]. But since the sum of these two functionals is a constant $H_Y(S) = H(Q) < \infty$, both of them must be continuous. The continuity of $H_{X|Y}$ and $H_{X,Y}$ now follows from (5).

Now consider $\mathcal{C}(P, m)$. In [17] it is shown that $H(Y|X)$ and $I(X; Y)$ are continuous when the alphabet of Y is finite and fixed, which is what we have here. And since $H(X) = H(P)$ is fixed, $H(X|Y)$ and $H(X, Y)$ are also continuous (if $H(P) = \infty$ then they are infinite over the entire $\mathcal{C}(P, m)$, but we also take this to mean that they are continuous). \blacksquare

In fact, since $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$ are compact, Shannon information measures are *uniformly* continuous over these domains, for any P, Q with finite entropy, and $m \in \mathbb{N}$.

Combining the above results, we obtain the following.

Theorem 7: Let $P, Q \in \Gamma^{(1)}$ and $m \in \mathbb{N}$, and assume that Q has finite entropy. Then Shannon information measures attain their extreme values over $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$.

Proof: The claim follows from Theorems 3 and 6: Every continuous function attains its infimum and supremum over a compact set [31, Thm 4.16]. \blacksquare

This claim is obsolete for the maximum entropy distribution because it is easy to find it. Namely, $P \times Q = (p_i q_j)$ maximizes entropy over $\mathcal{C}(P, Q)$ as is easily seen from (6), and $P \times U_m$ is the maximizer over $\mathcal{C}(P, m)$, where U_m is the uniform distribution over $\{1, \dots, m\}$. But it is much harder to find the minimum entropy distribution, as we will show, and its existence is not obvious when the alphabets are unbounded.

The argument in the proof of Theorem 6 can easily be adapted to prove the following more general claim.

Theorem 8: Let $S_n, S \in \Gamma^{(2)}$ be bivariate probability distributions. If S_n converges to S ($\|S_n - S\|_1 \rightarrow 0$) in such a

way that $H_X(S_n) \rightarrow H_X(S)$ and $H_Y(S_n) \rightarrow H_Y(S)$, and if at least one of these marginal entropies is finite, then we must have:

$$\begin{aligned} H_{X,Y}(S_n) &\rightarrow H_{X,Y}(S), & I_{X;Y}(S_n) &\rightarrow I_{X;Y}(S) \\ H_{X|Y}(S_n) &\rightarrow H_{X|Y}(S), & H_{Y|X}(S_n) &\rightarrow H_{Y|X}(S). \end{aligned} \quad (35)$$

Proof: As in the proof of Theorem 6, we observe that $\liminf_{n \rightarrow \infty} H_{Y|X}(S_n) \geq H_{Y|X}(S)$ (by Lemma 2), and that $\liminf_{n \rightarrow \infty} I_{X;Y}(S_n) \geq I_{X;Y}(S)$ which follows from (34) and the fact that when $S_n \rightarrow S$, then also $P_n \times Q_n \rightarrow P \times Q$, where P_n, Q_n , and P, Q are the marginals of S_n and S , respectively. But since $H_Y(S_n) \rightarrow H_Y(S)$ by assumption, one sees from (33) that both of these inequalities must in fact be equalities. The remaining claims in (35) then follow from (5). ■

The previous claim establishes that if $\|S_n - S\|_1 \rightarrow 0$, then a sufficient condition for the convergence of joint entropy is the convergence of marginal entropies. It is also necessary, as the following theorem shows.

Theorem 9: Let S_n, S be bivariate probability distributions such that $\|S_n - S\|_1 \rightarrow 0$ and $H_{X,Y}(S_n) \rightarrow H_{X,Y}(S) < \infty$. Then $H_X(S_n) \rightarrow H_X(S)$ and $H_Y(S_n) \rightarrow H_Y(S)$, and consequently, all claims in (35) hold.

Proof: The claim follows from the identity $H_{X,Y}(S_n) = H_X(S_n) + H_{Y|X}(S_n)$, and the fact that H_X and $H_{Y|X}$ are both lower semi-continuous. ■

C. (Dis)continuity of Rényi entropy

Rényi entropy H_α is known to be a continuous functional for $\alpha > 1$ (see, e.g., [24]) and it of course remains continuous over $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$. Therefore, it is also bounded and attains its extrema over these domains. It is, however, in general discontinuous for $\alpha \in [0, 1]$, and its behavior over $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$ needs to be examined separately. The case $\alpha = 1$ (Shannon entropy) has been settled in the previous subsection, so in the following we assume that $\alpha \in [0, 1)$.

Theorem 10: H_α is continuous over $\mathcal{C}(P, m)$, for any $\alpha > 0$. For $\alpha = 0$ it is discontinuous for any $m \geq 2$.

Proof: Let $0 < \alpha < 1$. If $H_\alpha(P) = \infty$, then $H_\alpha(S) = \infty$ for any $S \in \mathcal{C}(P, m)$ and there is nothing to prove, so assume that $H_\alpha(P) < \infty$. Let S_n be a sequence of bivariate distributions converging to S , and observe:

$$\sum_{i,j} S_n(i, j)^\alpha. \quad (36)$$

Since $S_n(i, j) \leq P(i)$ and $\sum_{i=1}^\infty \sum_{j=1}^m P(i)^\alpha = m \sum_{i=1}^\infty P(i)^\alpha < \infty$ by assumption, it follows from the Weierstrass criterion [31, Thm 7.10] that the series (36) converges uniformly (in n) and therefore:

$$\begin{aligned} \lim_{n \rightarrow \infty} \sum_{i,j} S_n(i, j)^\alpha &= \sum_{i,j} \lim_{n \rightarrow \infty} S_n(i, j)^\alpha \\ &= \sum_{i,j} S(i, j)^\alpha \end{aligned} \quad (37)$$

which gives $H_\alpha(S_n) \rightarrow H_\alpha(S)$.

As for the case $\alpha = 0$ it is easy to exhibit a sequence $S_n \rightarrow S$ such that the supports of S_n strictly contain the support of

S , i.e., $|S_n| > |S|$, implying that $\lim_{n \rightarrow \infty} H_0(S_n) > H_0(S)$. The case $m = 1$ is uninteresting because $\mathcal{C}(P, 1) = \{P\}$. ■

Unfortunately, continuity over $\mathcal{C}(P, Q)$ fails in general, as we discuss next.

Theorem 11: For any $\alpha \in (0, 1)$ there exist distributions P, Q with $H_\alpha(P) < \infty$ and $H_\alpha(Q) < \infty$, such that H_α is unbounded over $\mathcal{C}(P, Q)$.

Proof: Let $P = Q = (p_i)$ and assume that the p_i 's are monotonically nonincreasing. Define S_n with $S_n(i, j) = \frac{p_n}{n^r} + \varepsilon_{i,j}$ for $i, j \in \{1, \dots, n\}$, where $\varepsilon_{i,j} > 0$ are chosen to obtain the correct marginals and $r > 1$, and $S_n(i, j) = p_i \delta_{i,j}$ otherwise, where $\delta_{i,j}$ is the Kronecker's delta. Then $S_n \in \mathcal{C}(P, Q)$, and

$$\sum_{i,j} S_n(i, j)^\alpha \geq \sum_{i=1}^n \sum_{j=1}^n \left(\frac{p_n}{n^r}\right)^\alpha = n^{2-r\alpha} p_n^\alpha \quad (38)$$

Now, if p_n decreases to zero slowly enough, the previous expression will tend to ∞ when $n \rightarrow \infty$ for appropriately chosen r . For example, let $p_n \sim n^{-\beta}$, $\beta > 1$. Then whenever $2 - r\alpha - \beta\alpha > 0$, i.e., $r + \beta < 2\alpha^{-1}$, we will have $\lim_{n \rightarrow \infty} H_\alpha(S_n) = \infty$. Furthermore, if $\beta\alpha > 1$, then $H_\alpha(P) < \infty$. Therefore, for a given $\alpha \in (0, 1)$, we have found distributions P and Q with finite entropy of order α , such that H_α is unbounded over $\mathcal{C}(P, Q)$. ■

It is known that Rényi entropy H_α satisfies $H_\alpha(X, Y) \leq H_\alpha(X) + H_\alpha(Y)$ only for $\alpha = 0$ and $\alpha = 1$. Such an upper bound does not hold for $\alpha \in (0, 1)$, and, in fact, no upper bound on $H_\alpha(X, Y)$ in terms of $H_\alpha(X)$ and $H_\alpha(Y)$ can exist, as Theorem 11 shows.

Corollary 12: For any $\alpha \in (0, 1)$ there exist distributions P and Q such that H_α is discontinuous at every point of $\mathcal{C}(P, Q)$.

Proof: Let P and Q be such that H_α is unbounded over $\mathcal{C}(P, Q)$. Let S be an arbitrary distribution from $\mathcal{C}(P, Q)$. It is enough to show that H_α remains unbounded in any neighborhood of S . Let $M > 0$ be an arbitrary number, and $\epsilon \in (0, 1)$. We can find $T \in \mathcal{C}(P, Q)$ with $H_\alpha(T)$ as large as desired, so assume that $\sum_{i,j} t_{i,j}^\alpha \geq M/\epsilon$. Observe the distribution $(1 - \epsilon)S + \epsilon T$. It is in 2ϵ -neighborhood of S since $\|S - ((1 - \epsilon)S + \epsilon T)\|_1 = \epsilon\|S - T\|_1 \leq 2\epsilon$. Also, since the function x^α is concave for $\alpha < 1$, we get:

$$\begin{aligned} \sum_{i,j} ((1 - \epsilon)s_{i,j} + \epsilon t_{i,j})^\alpha &\geq \\ (1 - \epsilon) \sum_{i,j} s_{i,j}^\alpha + \epsilon \sum_{i,j} t_{i,j}^\alpha &\geq M, \end{aligned} \quad (39)$$

which completes the proof. ■

The case of $\alpha = 0$ (Hartley entropy) remains; the proof of the following result is straightforward.

Theorem 13: H_0 is discontinuous over $\mathcal{C}(P, Q)$, for any distributions P and Q with supports of size at least two.

Note that, unlike for the Shannon information measures, we cannot claim in general that H_α attains its supremum over $\mathcal{C}(P, Q)$, for $\alpha < 1$. However, infimum is attained, i.e., *minimum α -entropy coupling always exists*, because Rényi entropy is lower semi-continuous [24], and any such function must attain its infimum over a compact set [19].

We next prove that, although H_α is discontinuous for some P and Q , the continuity still holds for a wide class of marginal distributions.

Theorem 14: If $\sum_{i,j} \min\{p_i, q_j\}^\alpha < \infty$, then H_α is continuous over $\mathcal{C}(P, Q)$, for any $\alpha > 0$. For $P = Q = (p_i)$, with p_i 's nonincreasing, this condition reduces to $\sum_i i \cdot p_i^\alpha < \infty$.

Proof: Let $S_n \rightarrow S$, where $S_n, S \in \mathcal{C}(P, Q)$. Since, over $\mathcal{C}(P, Q)$, $S_n(i, j) \leq \min\{p_i, q_j\}$ and by assumption $\sum_{i,j} \min\{p_i, q_j\}^\alpha < \infty$, we can apply the Weierstrass criterion to conclude that $\sum_{i,j} S_n(i, j)^\alpha$ converges uniformly in n and therefore that $H_\alpha(S_n) \rightarrow H_\alpha(S)$.

Now let $P = Q$ and assume that the p_i 's are monotonically nonincreasing. Then $\min\{p_i, p_j\} = p_{\max\{i,j\}}$, i.e.,

$$(\min\{p_i, p_j\}) = \begin{pmatrix} p_1 & p_2 & p_3 & \cdots \\ p_2 & p_2 & p_3 & \cdots \\ p_3 & p_3 & p_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (40)$$

By observing the elements above (and including) the diagonal, it follows that:

$$\sum_i i \cdot p_i^\alpha \leq \sum_{i,j} \min\{p_i, p_j\}^\alpha \leq 2 \sum_i i \cdot p_i^\alpha, \quad (41)$$

and hence the condition $\sum_i i \cdot p_i^\alpha < \infty$ is equivalent to $\sum_{i,j} \min\{p_i, p_j\}^\alpha < \infty$. ■

Finally, let us prove a result analogous to Theorem 9.

Theorem 15: Let S_n, S be bivariate probability distributions such that $\|S_n - S\|_1 \rightarrow 0$ and $H_\alpha(S_n) \rightarrow H_\alpha(S) < \infty$. Let P_n, Q_n be the marginals of S_n , and P, Q the marginals of S . Then $H_\alpha(P_n) \rightarrow H_\alpha(P)$ and $H_\alpha(Q_n) \rightarrow H_\alpha(Q)$.

Proof: If $\|S_n - S\|_1 \rightarrow 0$, then of course $\|P_n - P\|_1 \rightarrow 0$ and $\|Q_n - Q\|_1 \rightarrow 0$. Write:

$$\begin{aligned} \sum_{i,j} S_n(i, j)^\alpha &= \sum_i P_n(i)^\alpha + \\ &\quad \sum_i \left(\sum_j S_n(i, j)^\alpha - P_n(i)^\alpha \right) \end{aligned} \quad (42)$$

We are interested in showing that the first term on the right-hand side converges to $\sum_i P(i)^\alpha$, which is equivalent to saying that $H_\alpha(P_n) \rightarrow H_\alpha(P)$. Observe that this term is lower semi-continuous by Lemma 2, meaning that

$$\liminf_{n \rightarrow \infty} \sum_i P_n(i)^\alpha \geq \sum_i P(i)^\alpha, \quad (43)$$

The second term on the right-hand side of (42) is also lower semi-continuous for the same reason, namely:

$$\sum_j S_n(i, j)^\alpha - P_n(i)^\alpha \geq 0 \quad (44)$$

because the function x^α is subadditive, and

$$\lim_{n \rightarrow \infty} \left(\sum_j S_n(i, j)^\alpha - P_n(i)^\alpha \right) = \sum_j S(i, j)^\alpha - P(i)^\alpha, \quad (45)$$

because $H_\alpha(S_n) \rightarrow H_\alpha(S)$. Therefore,

$$\begin{aligned} \liminf_{n \rightarrow \infty} \sum_i \left(\sum_j S_n(i, j)^\alpha - P_n(i)^\alpha \right) &\geq \\ \sum_i \left(\sum_j S(i, j)^\alpha - P(i)^\alpha \right), \end{aligned} \quad (46)$$

or, since $\sum_{i,j} S_n(i, j)^\alpha \rightarrow \sum_{i,j} S(i, j)^\alpha$,

$$\limsup_{n \rightarrow \infty} \sum_i P_n(i)^\alpha \leq \sum_i P(i)^\alpha. \quad (47)$$

Now (43) and (47) give $H_\alpha(P_n) \rightarrow H_\alpha(P)$, and $H_\alpha(Q_n) \rightarrow H_\alpha(Q)$ follows by symmetry. ■

Note that the opposite implication does not hold for any $\alpha \in [0, 1)$, as Corollary 12 shows. Namely, if $\|S_n - S\|_1 \rightarrow 0$, convergence of the marginal entropies ($H_\alpha(P_n) \rightarrow H_\alpha(P)$ and $H_\alpha(Q_n) \rightarrow H_\alpha(Q)$) does not imply convergence of the joint entropy $H_\alpha(S_n) \rightarrow H_\alpha(S)$.

V. ENTROPY METRICS

Apart from many of their other uses, couplings are very convenient for defining metrics on the space of probability distributions. There are many interesting metrics defined via so-called ‘‘optimal’’ couplings. We first illustrate this point using one familiar example, and then define new information-theoretic metrics based on the minimum entropy coupling.

Given two probability distributions P and Q , one could measure the ‘‘distance’’ between them as follows. Consider all possible random pairs (X, Y) with marginal distributions P and Q . Then define some measure of dissimilarity of X and Y , for example $\mathbb{P}(X \neq Y)$, and minimize it over all such couplings (minimization is necessary for the triangle inequality to hold). Indeed, this example yields the well-known total variation distance [25]:

$$d_V(P, Q) = \inf_{\mathcal{C}(P, Q)} \mathbb{P}(X \neq Y), \quad (48)$$

where the infimum is taken over all joint distributions of the random vector (X, Y) with marginals P and Q . Notice that a minimizing distribution (called a *maximal coupling*, see, e.g., [33]) in (48) is ‘‘easy’’ to find because $\mathbb{P}(X \neq Y)$ is a linear functional in the joint distribution of (X, Y) . For the same reason, $d_V(P, Q)$ is easy to compute, but this is also clear from the identity [25]:

$$d_V(P, Q) = \frac{1}{2} \sum_i |p_i - q_i|. \quad (49)$$

Now let us define some information-theoretic distances in a similar manner. Let (X, Y) be a random pair with joint distribution S and marginal distributions P and Q . The total information contained in these random variables is $H(X, Y)$, while the information contained simultaneously in both of them (or the information they contain about each other) is measured by $I(X; Y)$. One is then tempted to take as a

measure of their dissimilarity²:

$$\begin{aligned}\Delta_1(X, Y) &\equiv \Delta_1(S) = H(X, Y) - I(X; Y) \\ &= H(X|Y) + H(Y|X).\end{aligned}\quad (50)$$

Indeed, this quantity (introduced by Shannon [34], and usually referred to as the *entropy metric* [10]) satisfies the properties of a pseudometric [10]. In a similar way one can show that the following is also a pseudometric:

$$\Delta_\infty(X, Y) \equiv \Delta_\infty(S) = \max\{H(X|Y), H(Y|X)\}, \quad (51)$$

as are the normalized variants of Δ_1 and Δ_∞ [8]. These pseudometrics have found numerous applications (see for example [40]) and have also been considered in an algorithmic setting [5].

Remark 1: Δ_1 is a pseudometric on the space of random variables over the same probability space. Namely, for Δ_1 to be defined, the joint distribution of (X, Y) must be given because joint entropy and mutual information are not defined otherwise. Equation (55) below defines the distance between random variables (more precisely, between their distributions) that does not depend on the joint distribution.

One can further generalize these definitions to obtain a family of pseudometrics. This generalization is akin to the familiar ℓ_p distances. Let

$$\Delta_p(X, Y) \equiv \Delta_p(S) = (H(X|Y)^p + H(Y|X)^p)^{\frac{1}{p}}, \quad (52)$$

for $p \geq 1$. Observe that $\lim_{p \rightarrow \infty} \Delta_p(X, Y) = \Delta_\infty(X, Y)$, justifying the notation.

Theorem 16: $\Delta_p(X, Y)$ satisfies the properties of a pseudometric, for all $p \in [1, \infty]$.

Proof: Nonnegativity and symmetry are clear, as is the fact that $\Delta_p(X, Y) = 0$ if (but not only if) $X = Y$ with probability one. The triangle inequality remains. Following the proof for Δ_1 from [10, Lemma 3.7], we first observe that $H(X|Y) \leq H(X|Z) + H(Z|Y)$, wherefrom:

$$\begin{aligned}\Delta_p(X, Y) &\leq \left((H(X|Z) + H(Z|Y))^p + \right. \\ &\quad \left. (H(Y|Z) + H(Z|X))^p \right)^{\frac{1}{p}}.\end{aligned}\quad (53)$$

Now apply the Minkowski inequality ($\|a + b\|_p \leq \|a\|_p + \|b\|_p$) to the vectors $a = (H(X|Z), H(Z|X))$ and $b = (H(Z|Y), H(Y|Z))$ to get:

$$\Delta_p(X, Y) \leq \Delta_p(X, Z) + \Delta_p(Z, Y), \quad (54)$$

which was to be shown. ■

Having defined measures of dissimilarity, we can now define the corresponding distances:

$$\underline{\Delta}_p(P, Q) = \inf_{S \in \mathcal{C}(P, Q)} \Delta_p(S). \quad (55)$$

The case $p = 1$ has also been analyzed in some detail in [37], motivated by the problem of optimal order reduction for stochastic processes.

Theorem 17: $\underline{\Delta}_p$ is a pseudometric on $\Gamma^{(1)}$, for any $p \in [1, \infty]$.

² Drawing a familiar information-theoretic Venn diagram [9] makes it clear that this is a measure of “dissimilarity” of two random variables.

Proof: Since Δ_p satisfies the properties of a pseudometric, we only need to show that these properties are preserved under the infimum. 1° Nonnegativity is clearly preserved, $\underline{\Delta}_p \geq 0$. 2° Symmetry is also preserved, $\underline{\Delta}_p(P, Q) = \underline{\Delta}_p(Q, P)$. 3° If $P = Q$ then $\underline{\Delta}_p(P, Q) = 0$. This is because $S = \text{diag}(P)$ (distribution with masses $p_i = q_i$ on the diagonal and zeroes elsewhere) belongs to $\mathcal{C}(P, Q)$ in this case, and for this distribution we have $H_{X|Y}(S) = H_{Y|X}(S) = 0$. 4° The triangle inequality is left. Let X, Y and Z be random variables with distributions P, Q and R , respectively, and let their joint distribution be specified. We know that $\Delta_p(X, Y) \leq \Delta_p(X, Z) + \Delta_p(Z, Y)$, and we have to prove that

$$\inf_{\mathcal{C}(P, Q)} \Delta_p(X, Y) \leq \inf_{\mathcal{C}(P, R)} \Delta_p(X, Z) + \inf_{\mathcal{C}(R, Q)} \Delta_p(Z, Y). \quad (56)$$

Since, from the above,

$$\begin{aligned}\inf_{\mathcal{C}(P, Q)} \Delta_p(X, Y) &= \inf_{\mathcal{C}(P, Q, R)} \Delta_p(X, Y) \\ &\leq \inf_{\mathcal{C}(P, Q, R)} \{ \Delta_p(X, Z) + \Delta_p(Z, Y) \}\end{aligned}\quad (57)$$

it suffices to show that

$$\begin{aligned}\inf_{\mathcal{C}(P, Q, R)} \{ \Delta_p(X, Z) + \Delta_p(Z, Y) \} &= \\ \inf_{\mathcal{C}(P, R)} \Delta_p(X, Z) + \inf_{\mathcal{C}(R, Q)} \Delta_p(Z, Y).\end{aligned}\quad (58)$$

($\mathcal{C}(P, Q, R)$ denotes the set of all three-dimensional distributions with one-dimensional marginals P, Q , and R , as the notation suggests.) Let $T \in \mathcal{C}(P, R)$ and $U \in \mathcal{C}(R, Q)$ be the optimizing distributions on the right-hand side (rhs) of (58). Observe that there must exist a joint distribution $W \in \mathcal{C}(P, Q, R)$ consistent with T and U (for example, take $w_{i,j,k} = t_{i,k}u_{k,j}/r_k$). Since the optimal value of the lhs is less than or equal to the value at W , we have shown that the lhs of (58) is less than or equal to the rhs. For the opposite inequality observe that the optimizing distribution on the lhs of (58) defines some two-dimensional marginals $T \in \mathcal{C}(P, R)$ and $U \in \mathcal{C}(R, Q)$, and the optimal value of the rhs must be less than or equal to its value at (T, U) . ■

Remark 2: If $\underline{\Delta}_p(P, Q) = 0$, then P and Q are a permutation of each other. This is easy to see because only in that case can one have $H_{X|Y}(S) = H_{Y|X}(S) = 0$, for some $S \in \mathcal{C}(P, Q)$. Therefore, if distributions are identified up to a permutation, then $\underline{\Delta}_p$ is a metric. In other words, if we think of distributions as unordered multisets of nonnegative numbers summing up to one, then $\underline{\Delta}_p$ is a metric on such a space.

Observe that the distribution defining $\underline{\Delta}_p(P, Q)$ is in fact the minimum entropy coupling. Thus minimum entropy coupling defines the distances $\underline{\Delta}_p$ on the space of probability distributions in the same way maximal coupling defines the total variation distance. However, there is a sharp difference in the computational complexity of finding these two couplings, as will be shown in the following section.

A. Some properties of entropy metrics

We first note that $\underline{\Delta}_p$ is a monotonically nonincreasing function of p . In the following, we shall mostly deal with $\underline{\Delta}_1$

and $\underline{\Delta}_\infty$, but most results concerning bounds and convergence can be extended to all $\underline{\Delta}_p$ based on this monotonicity property.

The metric $\underline{\Delta}_1$ gives an upper bound on the entropy difference $|H(P) - H(Q)|$. Namely, since

$$\begin{aligned} |H(X) - H(Y)| &= |H(X|Y) - H(Y|X)| \\ &\leq H(X|Y) + H(Y|X) \\ &= \Delta_1(X, Y), \end{aligned} \quad (59)$$

we conclude that:

$$|H(P) - H(Q)| \leq \underline{\Delta}_1(P, Q). \quad (60)$$

Therefore, entropy is continuous with respect to this pseudometric, i.e., $\underline{\Delta}_1(P_n, P) \rightarrow 0$ implies $H(P_n) \rightarrow H(P)$. Bounding the entropy difference is an important problem in various contexts and it has been studied extensively, see for example [18], [33]. In particular, [33] studies bounds on the entropy difference via maximal couplings, whereas (60) is obtained via minimum entropy couplings.

Another useful property, relating the entropy metric $\underline{\Delta}_1$ and the total variation distance, follows from Fano's inequality:

$$H(X|Y) \leq \mathbb{P}(X \neq Y) \log(|X| - 1) + h(\mathbb{P}(X \neq Y)), \quad (61)$$

where $|X|$ denotes the size of the support of X , and $h(x) = -x \log_2(x) - (1 - x) \log_2(1 - x)$, $x \in [0, 1]$, is the binary entropy function. Evaluating the rhs at the maximal coupling (the joint distribution which minimizes $\mathbb{P}(X \neq Y)$), and the lhs at the minimum entropy coupling, we obtain:

$$\underline{\Delta}_1(P, Q) \leq d_V(P, Q) \log(|P||Q|) + 2h(d_V(P, Q)). \quad (62)$$

This relation makes sense only when the alphabets (supports of P and Q) are finite. When the supports are also fixed it shows that $\underline{\Delta}_1$ is continuous with respect to d_V , i.e., that $d_V(P_n, P) \rightarrow 0$ implies $\underline{\Delta}_1(P_n, P) \rightarrow 0$. By Pinsker's inequality [10] then it follows that $\underline{\Delta}_1$ is also continuous with respect to information divergence, i.e., $D(P_n||P) \rightarrow 0$ implies $\underline{\Delta}_1(P_n, P) \rightarrow 0$.

The continuity of $\underline{\Delta}_1$ with respect to d_V fails in the case of infinite (or even finite, but unbounded) supports, which follows from (60) and the fact that entropy is a discontinuous functional with respect to the total variation distance. One can, however, claim the following.

Proposition 18: If $P_n \rightarrow P$ in the total variation distance, and $H(P_n) \rightarrow H(P) < \infty$, then $\underline{\Delta}_1(P_n, P) \rightarrow 0$.

Proof: In [16, Thm 17] it is shown that if $d_V(P_{X_n}, P_X) \rightarrow 0$ and $H(X_n) \rightarrow H(X) < \infty$, then $\mathbb{P}(X_n \neq Y_n) \rightarrow 0$ implies $H(X_n|Y_n) \rightarrow 0$, for any r.v.'s Y_n . Our claim then follows by specifying $P_{X_n} = P_n$, $P_X = P_{Y_n} = P$, and taking infimums on both sides of the implication. ■

We also note here that sharper bounds than the above can be obtained by using $\underline{\Delta}_\infty$ instead of $\underline{\Delta}_1$. For example:

$$|H(P) - H(Q)| \leq \underline{\Delta}_\infty(P, Q), \quad (63)$$

(with equality whenever the minimum entropy coupling of P and Q is such that Y is a function of X , or vice versa), and:

$$\underline{\Delta}_\infty(P, Q) \leq d_V(P, Q) \log \max\{|P|, |Q|\} + h(d_V(P, Q)). \quad (64)$$

We conclude this section with an interesting remark on the conditional entropy. First observe that the pseudometric Δ_p ($\underline{\Delta}_p$) can also be defined for random vectors (multivariate distributions). For example, $\Delta_1((X, Y), (Z))$ is well-defined by $H(X, Y|Z) + H(Z|X, Y)$. If the distributions of (X, Y) and Z are S and R , respectively, then minimizing the above expression over all tri-variate distributions with the corresponding marginals S and R would give $\underline{\Delta}_1(S, R)$. Furthermore, random vectors need not be disjoint. For example, we have:

$$\Delta_1((X), (X, Y)) = H(X|X, Y) + H(X, Y|X) = H(Y|X), \quad (65)$$

because the first summand is equal to zero. Therefore, the conditional entropy $H(Y|X)$ can be seen as the distance between the pair (X, Y) and the conditioning random variable X . If the distribution of (X, Y) is S , and the marginal distribution of X is P , then:

$$\underline{\Delta}_1(P, S) = H_{Y|X}(S), \quad (66)$$

because S is the only distribution consistent with these constraints. In fact, we have $\underline{\Delta}_p(P, S) = H_{Y|X}(S)$ for all $p \in [1, \infty]$. Therefore, the conditional entropy $H(Y|X)$ represents the distance between the joint distribution of (X, Y) and the marginal distribution of the conditioning random variable X .

VI. OPTIMIZATION PROBLEMS

In this final section we analyze some natural optimization problems associated with information measures over $\mathcal{C}(P, Q)$ and $\mathcal{C}(P, m)$, and establish their computational intractability. The proofs are not difficult, but they have a number of important consequences, as discussed in Section VI-C, and, furthermore, they give interesting information-theoretic interpretations of well-known problems in complexity theory, such as the SUBSET SUM and the PARTITION problems. Some closely related problems over $\mathcal{C}(P, Q)$, in the context of computing $\underline{\Delta}_1(P, Q)$, are also studied in [37].

A. Optimization over $\mathcal{C}(P, Q)$

Consider the following computational problem, called MINIMUM ENTROPY COUPLING: Given $P = (p_1, \dots, p_n)$ and $Q = (q_1, \dots, q_m)$ (with $p_i, q_j \in \mathbb{Q}$), find the minimum entropy coupling of P and Q . It is shown below that this problem is NP-hard. The proof relies on the following well-known NP-complete problem [12]:

Problem: SUBSET SUM

Instance: Positive integers d_1, \dots, d_n and s .

Question: Is there a $J \subseteq \{1, \dots, n\}$ such that $\sum_{j \in J} d_j = s$?

Theorem 19: MINIMUM ENTROPY COUPLING is NP-hard.

Proof: We shall demonstrate a reduction from the SUBSET SUM to the MINIMUM ENTROPY COUPLING. Let there be given an instance of the SUBSET SUM, i.e., a set of positive integers $s; d_1, \dots, d_n$, $n \geq 2$. Let $D = \sum_{i=1}^n d_i$, and let $p_i = d_i/D$, $q = s/D$ (assume that $s < D$, the problem otherwise being trivial). Denote $P = (p_1, \dots, p_n)$

and $Q = (q, 1 - q)$. The question we are trying to answer is whether there is a $J \subseteq \{1, \dots, n\}$ such that $\sum_{j \in J} d_j = s$, i.e., such that $\sum_{j \in J} p_j = q$. Observe that this happens if and only if there is a matrix S with row sums $P = (p_1, \dots, p_n)$ and column sums $Q = (q, 1 - q)$, which has exactly one nonzero entry in every row (or, in probabilistic language, a distribution $S \in \mathcal{C}(P, Q)$ such that Y deterministically depends on X). We know that in this case, and only in this case, the entropy of S would be equal to $H(P)$ [9], which is by (6) a lower bound on entropy over $\mathcal{C}(P, Q)$. In other words, if such a distribution exists, it must be the minimum entropy coupling. Therefore, if we could find the minimum entropy coupling, we could easily decide whether it has one nonzero entry in every row, thereby solving the given instance of the SUBSET SUM. ■

Now from (5) we conclude that the problems of minimization of the conditional entropies and maximization of the mutual information over $\mathcal{C}(P, Q)$ are also NP-hard. Furthermore, in the same way as above one can define the problem MINIMUM α -ENTROPY COUPLING, for any $\alpha \geq 0$, and establish its NP-hardness. Note that the reverse problems over $\mathcal{C}(P, Q)$, entropy maximization for example, are trivial for Shannon information measures. In the case of Rényi entropy the problem is in general not trivial, but it can be solved by standard convex optimization methods.

It would be interesting to determine whether the MINIMUM ENTROPY COUPLING belongs to FNP³, but this appears to be quite difficult. Namely, given the optimal solution, it is not obvious how to verify (in polynomial time) that it is indeed optimal. A similar situation arises with the decision version of this problem: Given P and Q and a threshold h , is there a distribution $S \in \mathcal{C}(P, Q)$ with entropy $H(S) \leq h$? Whether this problem belongs to NP is another interesting question (which we shall not be able to answer here). The trouble with these computational problems is that \mathbb{R} -valued functions are involved. Verifying, for example, that $H(S) \leq h$ might not be computationally trivial as it might seem because the numbers involved are in general irrational. We shall not go into these details further; we mention instead one closely related problem which has been studied in the literature:

Problem: Sqrt SUM

Instance: Positive integers d_1, \dots, d_n , and k .

Question: Decide whether $\sum_{i=1}^n \sqrt{d_i} \leq k$?

This problem, though “conceptually simple” and bearing certain resemblance with the above decision version of the entropy minimization problem, is not known to be solvable in NP [11] (it is solvable in PSPACE).

B. Optimization over $\mathcal{C}(P, m)$

Minimization of the joint entropy $H(X, Y)$ over $\mathcal{C}(P, m)$ is trivial. The reason is that $H(X, Y) \geq H(P)$ with equality iff Y deterministically depends on X , and so the solution is *any* joint distribution having at most one nonzero entry in each row (the same is true for H_α , $\alpha \geq 0$). Since $H(X)$ is fixed, this also minimizes the conditional entropy $H(Y|X)$. The other

two optimization problems considered so far, minimization of $H(X|Y)$ and maximization of $I(X; Y)$, are still equivalent because $I(X; Y) = H(X) - H(X|Y)$, but they turn out to be much harder. Therefore, in the following we shall consider only the maximization of $I(X; Y)$.

Let OPTIMAL CHANNEL be the following computational problem: Given $P = (p_1, \dots, p_n)$ and m (with $p_i \in \mathbb{Q}, m \in \mathbb{N}$), find the distribution $S \in \mathcal{C}(P, m)$ which maximizes the mutual information. This problem is the reverse of the channel capacity in the sense that now the input distribution (the distribution of the source) is fixed, and the maximization is over the conditional distributions. In other words, given a source, we are asking for the channel with a given number of outputs which has the largest mutual information. Since the mutual information is convex in the conditional distribution [9], this is again a convex maximization problem.

We describe next the well-known PARTITION (or NUMBER PARTITIONING) problem [12].

Problem: PARTITION

Instance: Positive integers d_1, \dots, d_n .

Question: Is there a partition of $\{d_1, \dots, d_n\}$ into two subsets with equal sums?

This is clearly a special case of the SUBSET SUM. It can be solved in pseudo-polynomial time by dynamic programming methods [12]. But the following closely related problem is much harder.

Problem: 3-PARTITION

Instance: Nonnegative integers d_1, \dots, d_{3m} and k with $k/4 < d_j < k/2$ and $\sum_j d_j = mk$.

Question: Is there a partition of $\{1, \dots, 3m\}$ into m subsets J_1, \dots, J_m (disjoint and covering $\{1, \dots, 3m\}$) such that $\sum_{j \in J_i} d_j$ are all equal? (The sums are necessarily k and every J_i has 3 elements.)

This problem is NP-complete in the strong sense [12], i.e., no pseudo-polynomial time algorithm for it exists unless $P=NP$.

Theorem 20: OPTIMAL CHANNEL is NP-hard.

Proof: We prove the claim by reducing 3-PARTITION to OPTIMAL CHANNEL. Let there be given an instance of the 3-PARTITION problem as described above, and let $p_i = d_i/D$, where $D = \sum_i d_i$. Deciding whether there exists a partition with described properties is equivalent to deciding whether there is a matrix $C \in \mathcal{C}(P, m)$ with the other marginal Q being uniform and C having at most one nonzero entry in every row (i.e., Y deterministically depending on X). This on the other hand happens if and only if there is a distribution $C \in \mathcal{C}(P, m)$ with mutual information equal to $H(Q) = \log m$, which is by (7) an upper bound on $I_{X;Y}$ over $\mathcal{C}(P, m)$. The distribution C would therefore necessarily be the maximizer of $I_{X;Y}$. To conclude, if we could solve the OPTIMAL CHANNEL problem with instance $(p_1, \dots, p_{3m}; m)$, we could easily decide whether the maximizer is such that it has at most one nonzero entry in every row, thereby solving the original instance of the 3-PARTITION problem. ■

Note that the problem remains NP-hard even when the number of channel outputs (m) is fixed in advance and is

³ The class FNP captures the complexity of function problems associated with decision problems in NP, see [28].

not a part of the input instance. For example, maximization of $I_{X;Y}$ over $\mathcal{C}(P, 2)$ is essentially equivalent to the PARTITION problem.

It is easy to see that the transformation in the proof of Theorem 20 is in fact *pseudo-polynomial* [12] which implies that OPTIMAL CHANNEL is strongly NP-hard and, unless $P=NP$, has no pseudo-polynomial time algorithm.

C. Some comments and generalizations

1) *Entropy minimization*: Entropy minimization, taken in the broadest sense, is a very important problem. Watanabe [38] has shown, for example, that many algorithms for clustering and pattern recognition can be characterized as suitably defined entropy minimization problems.

A much more familiar problem in information theory is that of entropy maximization. The so-called *Maximum entropy principle* formulated by Jaynes [20] states that, among all probability distributions satisfying certain constraints (expressing our knowledge about the system), one should pick the one with maximum entropy. It has been recognized by Jaynes, as well as many other researchers, that this choice gives the least biased, the most objective distribution consistent with the information one possesses about the system. Consequently, the problem of maximizing entropy under constraints has been thoroughly studied (see, e.g., [14], [21]). It has also been argued [22], [41], however, that minimum entropy distributions can be of as much interest as maximum entropy distributions. The MinMax information measure, for example, has been introduced in [22] as a measure of the amount of information contained in a given set of constraints, and it is based both on maximum and minimum entropy distributions.

One could formalize the problem of entropy minimization as follows: Given a polytope (by a system of inequalities with rational coefficients, say) in the set of probability distributions, find the distribution S^* which minimizes the entropy functional H . (If the coefficients are rational, then all the vertices are rational, i.e., have rational coordinates. Therefore, the minimum entropy distribution has finite description and is well-defined as an output of a computational problem.) This problem is strongly NP-hard and remains such over transportation polytopes, as established above.

2) *Rényi entropy minimization*: The problem of minimization of the Rényi entropies H_α over arbitrary polytopes is also strongly NP-hard, for any $\alpha \geq 0$. Note that, for $\alpha > 1$, this problem is equivalent to the maximization of the ℓ^α norm (see also [26], [6] for different proofs of the NP-hardness of norm maximization). Interestingly, however, the minimization of H_∞ is polynomial-time solvable; it is equivalent to the maximization of the ℓ^∞ norm [26]. For $\alpha < 1$, the minimization of Rényi entropy is equivalent to the minimization of ℓ^α (which is not a norm in the strict sense), a problem arising in compressed sensing [13].

Hence, as we have seen throughout this section, various problems from computational complexity theory can be reformulated as information-theoretic optimization problems. (Observe also the similarity of the Sqrt SUM and the minimization of Rényi entropy of order $1/2$.)

3) *Other information measures*: Maximization of mutual information is a very important problem in the general context. The so-called Maximum mutual information criterion has found many applications, e.g., for feature selection [2] and the design of classifiers [15]. Another familiar example is that of the capacity of a communication channel which is defined precisely as the maximum of the mutual information between the input and the output of a channel.

We have illustrated the general intractability of the problem of maximization of $I_{X;Y}$ by exhibiting two simple classes of polytopes over which the problem is strongly NP-hard (and we have argued that the same holds for the conditional entropy).

We also mention here one possible generalization of this problem – maximization of information divergence. Namely, since for $S \in \mathcal{C}(P, Q)$:

$$I_{X;Y}(S) = D(S||P \times Q), \quad (67)$$

one can naturally consider the more general problem of maximization of $D(S||T)$ when S belongs to some convex region and T is fixed. Formally, let INFORMATION DIVERGENCE MAXIMIZATION be the following computational problem: Given a rational convex polytope Π in the set of probability distributions, and a distribution T , find the distribution $S \in \Pi$ which maximizes $D(\cdot||T)$. This is again a convex maximization problem because $D(S||T)$ is strictly convex in S [10].

Corollary 21: INFORMATION DIVERGENCE MAXIMIZATION is NP-hard.

Note that the reverse problem, namely the minimization of information divergence, defines an information projection of T onto the region Π [10].

4) *Measures of statistical dependence*: We conclude this section with one more generalization of the problem of maximization of mutual information. Namely, this problem can also be seen as a statistical problem of expressing the largest possible dependence between two given random variables.

Consider the following statistical scenario. A system is described by two random variables (taking values in \mathbb{N}) whose joint distribution is unknown; only some constraints that it must obey are given. The set of all distributions satisfying these constraints is usually called a statistical model.

Example 1: Suppose we have two correlated information sources obtained by independent drawings from a discrete bivariate probability distribution, and suppose we only have access to individual streams of symbols (i.e., streams of symbols from either one of the sources, but not from both simultaneously) and can observe the relative frequencies of the symbols in each of the streams. We therefore “know” probability distributions of both sources (say P and Q), but we don’t know how correlated they are. Then the “model” for this joint source would be $\mathcal{C}(P, Q)$. In the absence of any additional information, we must assume that some $S \in \mathcal{C}(P, Q)$ is the “true” distribution of the source.

Given such a model, we may ask the following question: What is the largest possible dependence of the two random variables? How correlated can they possibly be? This question can be made precise once a dependence measure is specified, and this is done next.

A. Rényi [30] has formalized the notion of probabilistic dependence by presenting axioms which a “good” dependence measure ρ should satisfy. These axioms, adapted for discrete random variables, are listed below.

- (A) $\rho(X, Y)$ is defined for any two random variables X, Y , neither of which is constant with probability 1.
- (B) $0 \leq \rho(X, Y) \leq 1$.
- (C) $\rho(X, Y) = \rho(Y, X)$.
- (D) $\rho(X, Y) = 0$ iff X and Y are independent.
- (E) $\rho(X, Y) = 1$ iff $X = f(Y)$ or $Y = g(X)$.
- (F) If f and g are injective functions, then $\rho(f(X), g(Y)) = \rho(X, Y)$.

Actually, Rényi considered axiom (E) to be too restrictive and demanded only the “if part”. It has been argued subsequently [3], however, that this is a substantial weakening. We shall find it convenient to consider the stronger axiom given above. As an example of a good measure of dependence, one could take precisely the mutual information; its normalized variant $I(X; Y) / \min\{H(X), H(Y)\}$ satisfies all the above axioms.

We can now formalize the question asked above. Namely, let MAXIMAL ρ -DEPENDENCE be the following problem: Given two probability distributions $P = (p_1, \dots, p_n)$ and $Q = (q_1, \dots, q_m)$, find the distribution $S \in \mathcal{C}(P, Q)$ which maximizes ρ . The proof of the following claim is identical to the one given for mutual information (entropy) in Section VI-A and we shall therefore omit it.

Theorem 22: Let ρ be a measure of dependence satisfying Rényi’s axioms. Then MAXIMAL ρ -DEPENDENCE is NP-hard.

The intractability of the problem over more general statistical models is now a simple consequence.

REFERENCES

- [1] J. Aczél and Z. Daróczy, *On Measures of Information and Their Characterization*, New York: Academic, 1975.
- [2] R. Battiti, “Using Mutual Information for Selecting Features in Supervised Neural Net Learning,” *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, Jul. 1994.
- [3] C. B. Bell, “Mutual Information and Maximal Correlation as Measures of Dependence,” *The Annals of Mathematical Statistics*, vol. 33, no. 2, pp. 587–595, Jun. 1962.
- [4] M. Ben-Bassat and J. Raviv, “Rényi’s Entropy and the Probability of Error,” *IEEE Trans. Inf. Theory*, vol. 24, no. 3, pp. 324–331, May 1978.
- [5] C. H. Bennett, P. Gács, M. Li, P. M. B. Vitányi, and W. H. Zurek, “Information Distance,” *IEEE Trans. Inf. Theory*, vol. 44, no. 4, pp. 1407–1423, Jul. 1998.
- [6] H. L. Bodlaender, P. Gritzmann, V. Klee, and J. Van Leeuwen, “Computational Complexity of Norm-Maximization,” *Combinatorica*, vol. 10, no. 2, pp. 203–225, Jun. 1990.
- [7] R. A. Brualdi, *Combinatorial Matrix Classes*, Cambridge University Press, 2006.
- [8] J.-F. Coeurjolly, R. Drouilhet, and J.-F. Robineau, “Normalized Information-Based Divergences,” *Probl. Inf. Transm.*, vol. 43, no. 3, pp. 167189, Sept. 2007.
- [9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley-Interscience, John Wiley and Sons, Inc., 2006.
- [10] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, Inc., 1981.
- [11] K. Etessami and M. Yannakakis, “On the Complexity of Nash Equilibria and Other Fixed Points,” *SIAM J. Comput.*, vol. 39, no. 6, pp. 2531–2597, 2010.
- [12] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*, W. H. Freeman and Co., 1979.
- [13] Dongdong Ge, Xiaoye Jiang, and Yinyu Ye, “A note on the complexity of L_p minimization,” *Math. Program., Ser. B*, vol. 129, pp. 285–299, Oct. 2011.
- [14] P. Harremoës and F. Topsøe, “Maximum entropy fundamentals,” *Entropy*, vol. 3, no. 3, pp. 191–226, Sept. 2001.
- [15] X. He, L. Deng, and W. Chou, “Discriminative Learning in Sequential Pattern Recognition,” *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 14–36, Sept. 2008.
- [16] S.-W. Ho and S. Verdú, “On the Interplay Between Conditional Entropy and Error Probability,” *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 5930–5942, Dec. 2010.
- [17] S.-W. Ho and R. W. Yeung, “On the Discontinuity of the Shannon Information Measures,” *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5362–5374, Dec. 2009.
- [18] S.-W. Ho and R. W. Yeung, “The Interplay between Entropy and Variational Distance,” *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 5906–5929, Dec. 2010.
- [19] R. B. Holmes, *Geometric Functional Analysis and Its Applications*, Springer-Verlag New York Inc., 1975.
- [20] E. T. Jaynes, “Information Theory and Statistical Mechanics,” *The Physical Review*, vol. 106, no. 4, pp. 620–630, May 1957.
- [21] J. N. Kapur, *Maximum-Entropy Models in Science and Engineering*, New Delhi, India: Wiley, 1989.
- [22] J. N. Kapur, G. Baciú, and H. K. Kesavan, “The MinMax Information Measure,” *Int. J. Syst. Sci.*, vol. 26, pp. 1–12, 1995.
- [23] M. Kovačević, I. Stanojević, and V. Šenk, “On the Hardness of Entropy Minimization and Related Problems,” in *Proc. IEEE Information Theory Workshop (ITW)*, Lausanne, Switzerland, 2012, pp. 512–516.
- [24] M. Kovačević, I. Stanojević, and V. Šenk, “Some Properties of Rényi Entropy over Countably Infinite Alphabets,” *Probl. Inf. Transm.*, accepted for publication, available at arXiv:1106.5130v2.
- [25] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times*, American Mathematical Society, 2008.
- [26] O. L. Mangasarian and T. H. Shiau, “A variable-complexity norm maximization problem,” *SIAM Journal on Algebraic and Discrete Methods*, vol. 7, no. 3, pp. 455–461, July 1986.
- [27] J. Munkres, *Topology*, 2nd ed., Prentice Hall Inc., 2000.
- [28] C. H. Papadimitriou, *Computational Complexity*, Addison-Wesley Publishing Company, Reading, MA, 1994.
- [29] A. Rényi, “On measures of entropy and information,” in *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.*, 1961, vol. I, Berkeley, CA: Univ. California Press, pp. 547–561.
- [30] A. Rényi, “On Measures of Dependence,” *Acta Math. Acad. Sci. Hungar.*, vol. 10, no. 3–4, pp. 441–451, Sept. 1959.
- [31] W. Rudin, *Principles of mathematical analysis*, 3rd ed., International Series in Pure and Applied Mathematics, McGraw-Hill Book Co., 1976.
- [32] L. Rüschendorf, B. Schweizer, and M. D. Taylor (Editors), *Distributions with Fixed Marginals and Related Topics*, Lecture Notes - Monograph Series, Institute of Mathematical Statistics, 1996.
- [33] I. Sason, “Entropy Bounds for Discrete Random Variables via Coupling,” *submitted to IEEE Trans. Inf. Theory*, available at arXiv:1209.5259v3.
- [34] C. E. Shannon, “Some Topics in Information Theory,” in *Proc. Int. Cong. Math.*, vol. 2, 262, 1950.
- [35] H. Thorison, *Coupling, Stationarity, and Regeneration*, Springer, 2000.
- [36] F. Topsøe, “Basic Concepts, Identities and Inequalities – the Toolkit of Information Theory,” *Entropy*, vol. 3, no. 3, pp. 162–190, Sept. 2001.
- [37] M. Vidyasagar, “A Metric Between Probability Distributions on Finite Sets of Different Cardinalities and Applications to Order Reduction,” *IEEE Trans. Aut. Control*, vol. 57, no. 10, pp. 2464–2477, Oct. 2012.
- [38] S. Watanabe, “Pattern recognition as a quest for minimum entropy,” *Pattern Recognit.*, vol. 13, no. 5, pp. 381–387, 1981.
- [39] A. Wehrl, “General properties of entropy,” *Rev. Mod. Phys.*, vol. 50, no. 2, pp. 221–260, Apr. 1978.
- [40] Y. Y. Yao, “Information-theoretic measures for knowledge discovery and data mining,” in *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, pp. 115–136, 2003.
- [41] L. Yuan and H. K. Kesavan, “Minimum entropy and information measure,” *IEEE Transactions on Systems, Man and Cybernetics - Part C*, vol. 28, pp. 488–491, 1998.